EVALUATION DRIVEN RESEARCH: The Foundation of the TIPSTER Text Program

Dr. John D. Prange
Department of Defense
jdprang@afterlife.ncsc.mil

INTRODUCTION:

I have been fortunate to have had the opportunity to be associated with the TIPSTER Text Program since its inception in 1989. Preliminary discussions among government researchers who were interested in establishing a major, new inter-agency text handling, processing, and exploitation program began in the Summer of that year and continued in earnest during the months that followed. Most of our frequent day-long planning meetings during the first year of our TIPSTER Program planning were held at DARPA headquarters in Arlington, VA and were chaired by the Program Manger of DARPA's Speech and Text R&D efforts.

There were clearly two different sets of experience and expertise present during these meetings.

The DARPA Program Manager was a strong proponent and advocate of an Evaluation Driven Research Paradigm that he was following in the Speech component of his R&D program. And even though this Program Manager had spent only slightly more than one year at DARPA, he clearly understood how DARPA established, funded and managed new R&D Programs.

The rest of us around the table had little or no previous exposure to either the details of an Evaluation Driven Research Paradigm or to the inner workings of DARPA programs. What we brought to the table were strong credentials and experience in artificial intelligence, natural language processing and computational linguistics. We also came with numerous challenging problems to be solved along with an understanding and appreciation of the text handling, processing, and exploitation needs of our individual agency's analysts and linguists.

The TIPSTER Text Program was born out of the best combination of these two camps. Since its creation in 1989, TIPSTER has developed, grown and evolved into its current role as a major driving force within both the Information Retrieval and Information Extraction R&D communities. TIPSTER has just completed its second, two-year Phase and is poised to begin Phase III, a three-year effort this coming October. Rather than running out of steam, TIPSTER has continued to pick up momentum and to broaden its area of interest and coverage as it proceeded through Phases I and II and now heads into Phase III.

Why has this happened? Looking back now from the perspective and vantage point of seven years of rich history, it is very clear to me that those of us who participated in these early, formative TIPSTER Text Program discussions collectively laid a very solid foundation. That foundation was built, "TIPSTER style", out of an Evaluation Driven Paradigm, heavily borrowed from DARPA's Speech R&D Program and it has continued to grow and evolve over the past seven years in a "TIPSTER unique" way.

EVALUATION DRIVEN RESEARCH: *What is it and what does it take to make it work?*

In the late 1980's the Speech and Text Technology Program at the Defense Advance Research Project Agency (DARPA) was heavily weighted and dominated by its Speech component. Those of us in the Government who were attempting to focus increasing attention and resources on text handling, processing and exploitation problems being encountered by our Agency's linguists and analysts looked with envy at the Speech component of this DARPA program. Increasingly, we wondered what it was that gave the DARPA Speech R&D Program its focus, its momentum, its continuity, its longevity, and most importantly its ability to dramatically move forward the state-of-the-art in their very challenging technical field.

While there were surely many reasons for the on-going success of DARPA's Speech R&D

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 1996		**00-00-1996 to 00-00-1996**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Evaluation Driven Research: The Foundation of the TIPSTER Text Program	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Department of Defense,Washington,DC,20301**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop held at Vienna, Virginia, May 6-8, 1996. Sponsored by the Defense Advanced Research Projects Agency.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**10**	

Program, a single answer emerged to us as text researchers during our early TIPSTER discussions. From our vantage point, the continuing vitality of DARPA's Speech R&D Program could be directly attributed to its enthusiastic implementation of an "Evaluation Driven Research" paradigm.

During these TIPSTER Program planning meetings in late 1989 and early 1990, we took a closer look at this implementation by DARPA's Speech R&D Community with an eye towards attempting to duplicate and tailor, as needed, their "Evaluation Driven Research" paradigm in our new Text Processing R&D Program. We identified the following distinct components which we hoped to carry over to our new text program:

- A clearly defined final objective for the overall R&D program.
- A series of specific tasks which when successfully accomplished would move the R&D community significantly closer to the program's final objective.
- An agreed upon and specifically tailored metric and evaluation methodology for periodically measuring progress towards accomplishing each of the chosen tasks.
- Sufficient quantities of training and testing data. Each data collection should be carefully selected, formatted, annotated, and otherwise prepared to directly support a specific task.
- A group of several (in fact, the more the merrier) leading-edge research institutions who are willing to:
  - ◊ Aggressively investigate solutions to each assigned task.
  - ◊ Periodically participate in formal evaluations of how well their systems are performing on the current task.
  - ◊ Openly discuss their successes and failures in the forum of a technical workshop attended by researchers from the other participating institutions and by interested government sponsors.
- A multi-year program budget with sufficient, programmed government funding to cover the cost of:
  - ◊ Obtaining and preparing the training and test data collections.
  - ◊ Fully funding the R&D activities of a core group of research institutions.
  - ◊ Providing strong encouragement and even some limited financial support to other non-core group research institutions to participate to the greatest extent possible in these periodic task evaluations and related open forums.
  - ◊ Conducting regularly scheduled formal evaluations of each task according to its agreed upon metric.
  - ◊ Sponsoring regularly scheduled open workshops to discuss and share results, approaches, and techniques.

On the surface, there appeared to us to be nothing revolutionary about these components and their description. We concluded that simply including them in a new R&D program would not, by themselves, guarantee its success. Rather we felt that the real key to the successful implementation of the "Evaluation Driven Research" paradigm lay in the careful and thoughtful selection of various program design choices and then in their actual execution. In particular several implementation considerations directly related to the paradigm components listed above seemed to be particularly important.

- Choosing an appropriate series of tasks.
  - ◊ Each task must be focused and clearly defined. If more than one task is being pursued simultaneously, there needs to be an overarching concept or framework into which these multiple tasks could meaningfully fit.
  - ◊ Each task must be technically challenging, and clearly a significant step beyond the current state-of-the-art. Each task must be far enough away so that no participant can reach it easily but not so far away that no one will make an attempt. The goal here is to spark the intellectual curiosity and interest of the world-class researchers in the field. Once you have attracted these pace setters and have enticed them to work on your task, then you have, in effect, indirectly moved many of the other institutions in this technical field who will independently take up this same task so that they can

stay up with this forward moving, state-of-the-art standard bearer.

- ◊ Another key consideration is to choose each task so that more than one technical approach can be followed and investigated. Clearly, stimulating a sense of friendly competition between participating research institutions within a single approach can produce improved results, but if this same sense of competition is established across multiple approaches, the net, positive result can be significantly multiplied. Belief in the "superiority" of a specific technical approach can be a tremendous motivator.
- ◊ And finally, the task must be stated and defined in such a way that it immediately lends itself to measurement and evaluation. In particular the best tasks are those whose rationale or case for action is stated in terms of the measurable amount of improvement which is being anticipated and sought.

- Choosing a metric and evaluation methodology. The strategy here is to choose a metric which is sufficiently close to the core research problem associated with each task that progress as measured by this metric will clearly imply that similar progress on the core technology which lies at the heart of the given task is also being made.
  - ◊ Choosing an easily accomplished yet inappropriate metric and evaluation methodology is clearly wasted energy. You can produce lots of data but yet produce few, if any, insights into either the effectiveness or efficiency of the underlying algorithm or in the progress being made on the actual, desired task.
  - ◊ Evaluation shines a bright light and attracts a lot of attention. We must understand the current task well enough to insure that our selected metrics will do the job they are intended to do; namely measuring the rate and degree of progress in achieving a satisfactory solution to the current task. Whatever we ultimately decide to evaluate is almost certainly guaranteed to receive significantly greater attention during the execution of this task. That's great if this is where attention needs to be applied, but devastating if the "real problem" to be solved lies elsewhere. This latter possibility is reminiscent of a story about a man who is walking home late at night and encounters a man, down on his hands and knees underneath a street light, carefully and systematically searching the ground with his hands. The first man asks the man on his knees what he is searching for. "I'm trying to find my lost watch", he replies. The first man willingly joins in the search. After several minutes have passed in unsuccessful search, the first man asks a second question, "Now, exactly where did you lose your watch?" The original searcher points across the street and replies, "Over there." "Then why are we searching here?" the first man responds. To which the original searcher replies, "The light is so much better here." People will naturally search longer and more completely where the light is strongest. Likewise, metrics and evaluation shine very bright lights. We must make sure that they illuminate those parts of the problem which deserve and will benefit from this special form of attention.
  - ◊ The selected metrics and evaluation methodology must simultaneously satisfy significantly different audiences. A high level view of the selected metrics must be simple enough that they can be easily understood by analyst end-users and by operational managers who are knowledgeable about the environment and domain into which these research results will eventually be applied but who are not necessarily technically savvy about the details of the underlying technology and algorithmic approaches being used. These analyst end-users and operational managers must be able to see the connection between the current task, the metrics being applied and the future impact that this emerging technology could have against their operational problems. Second, a subset of these metrics must allow the government sponsors to measure progress over time so that the government's return on investment can be tracked and

appropriate programmatic decisions reached. And finally, the full set of metrics must be sufficiently detailed and specific, so that the participating researchers can make adjustments in their algorithms and techniques based upon the insights that they have gained. Simultaneously satisfying these varied conditions and requirements can be difficult to achieve but it can be done and it is well worth the effort.

- Making available relevant training and testing data, in sufficient quantities and of appropriate quality. This is an area in which DARPA's Speech R&D Program has placed high value and importance and has clearly excelled in its execution. The key idea here is that the parameters associated with each collection of training and testing data must be carefully considered and selected prior to the beginning of each new R&D task. These parameters then became the guiding principles during the data preparation phase.
  - ◊ If the data does not already exist, then it must be created. Successful execution of the Evaluation Driven Research Paradigm is totally dependent upon the availability of high quality data in sufficient quantities which has been specifically prepared and formatted to support the current R&D tasks. While the creation of appropriate training and testing databases can be expensive and time consuming, these negatives must be accepted as part of the cost of doing business and must be factored into the budget and time schedule for that particular R&D task.
  - ◊ The optimal situation is one in which the data collection effort is 100% completed prior to the start of the associated research task. This may, however, be an unrealistic expectation. So when this is not practical, it is still technically possible for the data collection efforts to be done simultaneously with the task execution without significantly impacting the research efforts on the given task provided that the bare minimum data quantity and quality requirements are met on-time and that the parameters associated with the data preparation task remain constant throughout the entire effort. However, keeping this approach on track is easier said than done. The DARPA's Speech R&D Program has on more than one occasion used this simultaneous program and data preparation approach with mixed results. More on this subject later when we discuss this component in terms of the TIPSTER Program.
  - ◊ Happily, these specifically created training and testing databases can often be made available to other researchers and hence support other, initially unplanned R&D investigations. That is, the utility of these specially produced databases can be extended beyond their original purposes and yield important, unplanned side benefits.
  - ◊ The requisite preparation of data may be viewed as a mundane, routine, unexciting activity. But as previously stated, satisfactory completion of this data preparation task is absolutely essential. The degree of care and attention to detail that is applied during this activity directly translate into the quantity and quality of the R&D results produced by those investigations which utilize these data collections. There are no short cuts here.

- Fostering a cooperative, corporate program viewpoint among participating institutions and sponsoring government agencies.
  - ◊ The objective here is to make the program participants truly believe that each task to be solved takes precedence over who achieves the solution or by what method. While this may be overly idealistic, this objective's sentiment is exactly what is needed.
  - ◊ During the pursuit of each task, each participating institution will need to develop solutions to peripheral problems which are common to other technical approaches Every participant will need to input the same training and test data, to access similar collections of supporting information (*e.g.* lexicons, word lists, gazetteers), and to use functionally similar software tools (*e.g.* part of speech taggers, text annotation tools, segmentors). The participants should be

encouraged to equitably share these tools and data resources with other program participants. All stand to gain in the long run from such free and open exchanges.

◊ A great motivation for this viewpoint is the fact that continuing governmental funding support can best be secured by clearly demonstrating progress across a broad technological front. And this objective is easier to achieve in a cooperative, sharing environment, than in an isolated, proprietary one.

◊ Fostering and maintaining a cooperate and cooperative viewpoint applies to the team of government sponsors as well. It is hard enough for multiple offices within the same Agency is work together on the same program over an extended period of time, but this becomes much harder when you must factor in the cultural differences that will surely arises across several Agencies. When we looked closely at how DARPA's Speech R&D Program was managed we saw a fairly loose and unstructured confederation approach of interested government sponsors. By adopting this loose confederation, the Speech Program had avoided the need to confront the cultural differences across the sponsoring Agencies. On the other hand, as TIPSTER planners we hoped to establish a program which was equally planned, funded, managed, and executed by multiples Agencies. We were going to confront our Agencies' cultural differences head on. We knew that we had bitten off a lot.

EVALUATION DRIVEN RESEARCH: *TIPSTER Style*

So how well has TIPSTER adhered to the Evaluation Driven Research Paradigm as described in this preceding section? My assessment, in a phrase, is very well. Unfortunately a detailed response to this question is beyond the scope of this paper since the full answers to this question lies in the collective papers contained in the *Proceedings of the TIPSTER Text Program (Phase I)*, the Proceedings for each of the recent Message Understanding Conferences (MUC) and for each of the Text Retrieval Conferences (TREC) and the rest of this Proceedings for Phase II. So my objective for the remainder of this paper is to give a high level summary response to each paradigm component and maybe in the process to give a perspective with which you can read and interpret these individual papers.

Components of the Evaluation Driven Research Paradigm:

- A clearly defined final objective for the overall R&D program.

During the 1989-90 DARPA planning meetings a large number of important, yet diverse text handling, processing, and exploitation requirements surfaced. To make matters worse, each of these requirements took on many different forms when we took into account specific applications. Early on we opted to focus the TIPSTER Program on two core problems which seemed to be central to a large number of different operational problems. These two enabling technology areas are now well known and closely associated with the TIPSTER Program: Document Detection and Information Extraction. In Phase I the research goal was to significantly push the state-of-the-art in both fields using multiple, different technical approaches. In Phase II the research goals shifted. The main focus was now placed on investigating ways in which the two separate technology areas of document detection and information extraction could synergistically interact within a single, modular TIPSTER system architecture, on developing and deploying operational prototypes based upon the most promising TIPSTER algorithms, and on the continuing advancement of the overall performance of the best TIPSTER algorithms. In Phase III, we will add a third enabling technology area; text summarization while continuing to pursue natural extensions of these Phase II goals.

- A series of specific tasks which when successfully accomplished would move the R&D community significantly closer to the program's final objective.

The manner in which the TIPSTER Program has incorporated this component is most easily seen in the design of the multiple tasks that underwrote Phase I. Our evaluation of the pre-TIPSTER state-of-the-art in document detection systems concluded that there was:

◊ Heavy reliance on Boolean-logic key word systems

- ◊ Poor system performance
- ◊ Query construction required a system expert
- ◊ Bulk of the Information Retrieval (IR) research community efforts were directed at English documents and retrospective (or ad hoc) retrieval applications
- ◊ Many IR research systems were automatic, statistically-based systems whose performance on small, homogeneous research collections was promising but whose performance on real-world size collections was untested and unknown.

Similarly our evaluation of the pre-TIPSTER state-of-the-art in information extraction concluded that:

- ◊ Almost all work had been done in English
- ◊ The subject domains were focused primarily on military domains
- ◊ The input texts were typically highly structured and stylized
- ◊ The input text volumes were very small and system throughput was very slow
- ◊ Systems were designed to solve very specific applications. As a result system designs were highly "stove-piped". System portability was virtually non-existent.
- ◊ Systems failed "hard" when they encountered previously unseen vocabulary, linguistic structures, formats, etc.
- ◊ Practical applications were limited to highly constrained domains with high enough priority to warrant the development expense associated with a highly tailored system solution

In response to these conclusions, Phase I of TIPSTER established multiple, inter-related tasks. All participants were required to demonstrate language portability by performing the same basic tasks in both English and in Japanese and system robustness by successfully handling and processing text documents which contained ungrammatical usage, garbles, new words, and structures.

In addition the document detection participants were required to perform both routing and ad hoc retrieval tasks, to automatically convert detailed, lengthy natural language information need statements covering a wide range of topics into system specific queries without human intervention, to return relevant documents in priority order based upon the document's perceived degree of relevance, to highlight the most relevant passages within these retrieved documents, and to perform all of their tasks on large (now over 1 million documents and multiple gigabytes), heterogeneous, complex document collections.

Similarly the information extraction participants were additionally required to automatically locate, identify and standardize information contained in newspaper style documents within two distinct subject domains; the formation of business joint ventures and microelectronic chip fabrication. This entire extraction task was significantly more difficult than previous extraction tasks when measured along several dimensions (i.e. text corpus complexity, text corpus size, template fill complexity, and the overall nature of the task).

One of the most challenging information extraction tasks which was first articulated during Phase I (namely, system extensibility by analyst end-users) has still not been completely satisfied. Extraction systems are still best extended and modified by the system developers themselves or by individuals who have received significant training.

- An agreed upon and specifically tailored metric and evaluation methodology for periodically measuring progress towards accomplishing each of the chosen tasks.

Frequent formal metric-based evaluations have been a hallmark of the TIPSTER Text Program. The relevant evaluations are only highlighted in the following paragraphs. Each of these evaluations has been reported on in detail in either the *Proceedings of the TIPSTER Text Program (Phase I)*, in this Proceedings for Phase II or in the separately published Proceedings for the Message Understanding Conferences (MUC-3 to MUC-6) and Text Retrieval Conferences (TREC-1 to TREC-4). A reader wanting additional details is directed to one or more of these references.

During Phase I, all TIPSTER participants were formally evaluated shortly before the 12, 18, and 24 month Workshops.

In addition, the TIPSTER Text Program established close ties with the Message Understanding Conference (MUC) beginning with MUC-3. All of the TIPSTER Information Extraction contractors were required to participate in MUC-4

where the subject domain consisted of news reports on terrorism events. MUC-5 coincided with the TIPSTER Phase I 24-month evaluation and consisted of the same information extraction tasks that had been assigned to the Phase I participants (Formation of business joint ventures and microelectronic chip fabrication; each domain in two languages, English and Japanese). The non-TIPSTER MUC-5 participants could choose which of the 4 domain-language pairs they wished to be evaluated against. In November 1995, a redesigned MUC-6 has held in which each participant could choose to be evaluated in one or more of the following tasks; a named entity task, a template element task, a scenario template task, and a co-reference task. All four of these tasks were done using English source texts. In May 1996, TIPSTER sponsored a new information extraction evaluation program; the Multilingual Evaluation Task (MET). In MET, the participants performed the MUC-6 named entity task in one or more of the following foreign languages; Spanish, Chinese, and Japanese.

Early into Phase I of the TIPSTER Text Program, the decision was made to establish a companion evaluation program based initially on the TIPSTER Phase I document detection tasks. This companion evaluation program became known as the Text Retrieval Conference (TREC). To date, four TREC's have been held and the fifth is currently in progress. During TREC-1 to TREC-3, each participant was evaluated against both a routing task and an ad hoc retrieval task, each consisting of 50 test cases. Beginning with TREC-4, several additional specialty subtasks (referred to within TREC as Tracks) were added. These included a multiple database merging track, a confusion track to examine the effect of corrupted data, a multilingual track to examine retrieval of Spanish language documents, an interactive track, and a filtering track. These TREC Tracks are being continued in TREC-5. The major addition here is that the retrieval of Chinese language documents has been added to the multilingual track.

As part of TIPSTER Phase III, the TIPSTER R&D investigations will be expanded into the field of text summarization. Planning is already underway to determine an appropriate metric-based evaluation strategy for text summarization.

The impact of the TIPSTER Text Program metric-based evaluations can be readily seen from the single statistic that over 100 institutions have already participated in either a TIPSTER Text Program internal evaluation, or one or more of the MUC, MET, and TREC evaluation programs. In fact a significant majority of these institutions have participated at least twice and many have participated with even greater frequency.

- Sufficient quantities of training and testing data. Each data collection should be carefully selected, formatted, annotated, and otherwise prepared to directly support a specific task.

The thirteen different formal metric-based evaluations conducted variously under the banners of the TIPSTER Text Program Phase I (3), MUC (4), MET (1), and TREC (5) could not have been executed without sufficient quantities of training and testing data. The collection, annotation, tagging, and formatting of the base document collections along with the creation of the appropriate answer keys to support each separate evaluation program has been a costly, time consuming, human analyst intensive process. The bulk of these data preparation tasks were concentrated into Phase I, but additional data preparation efforts to support MUC, MET and TREC have continued, as needed, since the completion of Phase I in 1993. The performance of human analysts in completing their tasks has been routinely measured and have subsequently been used as a benchmark against which the performance of the information extraction and document detection algorithms can be compared.

As indicated earlier in this paper, the optimal situation is one in which the data collection effort is 100% completed prior to the start of the associated research task. This did not happen during TIPSTER Phase I. The collection, formatting and preparation of appropriate document databases and the creation of topic statements and pooled relevance judgments to support the document detection research tasks and of complex scenario templates, detailed fill rule descriptions, and appropriate answer keys to support the information extraction research task turned out to be a monumental undertaking. These data preparation tasks in both areas were several orders of magnitude greater than previous efforts. The TIPSTER government sponsors did not fully appreciate this fact until the data collection efforts were underway. We soon found ourselves in the situation where TIPSTER Phase I Program execution and data preparation were occurring simultaneously. It quickly proved very difficult, particularly on the information

extraction side, to maintain sufficient training and testing data throughput and at the same time, maintain high data consistency. While the job was eventually completed, it was only through the tireless and sometimes even heroic efforts of a small number of highly motivated and dedicated government researchers that this data preparation effort was brought to a successful conclusion in Phase I. To say the least, this is not a recommended mode of operation.

Again all of these TIPSTER data development activities have been previously reported on in the Proceedings associated with each of the evaluation programs identified earlier. The interested reader is directed to these sources for additional information and details.

- A group of several (in fact, the more the merrier) leading-edge research institutions who are willing to participate in a cooperative, corporate program.

The cooperativeness and corporateness of the TIPSTER Text Program participants has been repeatedly demonstrated in a wide variety of ways. A few examples are listed below to demonstrate the degree to which this statement has been played out.

◊ One participant in the Document Detection component of TIPSTER has participated in all three TIPSTER Phase I evaluations, in TREC-1 to TREC-4, and is currently participating in TREC-5. Likewise one participant in the Information Extraction component has participated in all three TIPSTER Phase I evaluations, MUC-3 to MUC-6, and MET. A number of other participants come close to matching these participation levels.

◊ Throughout the entire TIPSTER Text Program all of the contractors have willing shared data files and software modules with the other participants. This clearly allowed the collective program to cover more ground and to move forward faster.

◊ Since its beginning the TIPSTER Text Program has held technical workshops at 6 month intervals. The Phase II 24-month Workshop was the 10th such workshop. A portion of each workshop has been devoted to each contractor describing the technical details of their underlying algorithms and approaches, the results of their internally conducted evaluations and experiments, as well as their successes and failures on the TIPSTER sponsored formal evaluations. The openness of these presentations has always been highly commendable. To the degree that time permits, the same openness has been evident during each MUC, MET, and TREC. The importance of these forums and open discussions has been repeatedly demonstrated. A report outlining the details of successfully implemented techniques and approaches is made at one workshop by a single participant. Inevitably at the next workshop, reports are given by several other participants concerning how they were able to successfully and beneficially incorporate these new ideas into their own systems. In this way, a single success has been quickly multiplied.

◊ Establishing and maintaining a cooperative, corporate viewpoint among the program's external participants is made considerably easier if it is evident that there is a similar cooperative and corporate viewpoint being regularly demonstrated by the Government sponsors. Over the past seven years a unique bonding chemistry has developed among the large number of Government personnel who have had an active hand in the TIPSTER Program. Since October 1993 the introductory briefing of the TIPSTER Text Program has regularly been given as a joint briefing by Dr. Sarah Taylor of the Office of Research and Development and myself. This briefing has been frequently opened with the observation that "Multiple agencies have been working closely together on this Program since 1989. Why, in the process, we've even become friends." The line usually sparks a snicker or two, because those in the audience seem to know that previous joint programs between these Agencies have not always been so amicable. Almost from day one, there has been an underlying current of give and take, of teamwork, of consensus building. This atmosphere has proven to be quite contagious as new Government participants have joined the TIPSTER Program team and it has clearly rubbed off onto the other TIPSTER participants.

◊ In the Spring of 1994 the TIPSTER Text Program was nominated by the Community Management Staff as a "Reinvention Laboratory" in recognition of "its teamwork, its customer focus, and the fact that it has

broken down exiting bureaucratic barriers." Then in March 1996 Vice President Gore presented the National Performance Review Hammer Award to the TIPSTER Text Program in the reinvention of government. In his remarks, the Vice President lauded the TIPSTER Program's teamwork for spanning the Intelligence Community and partnering with the private sector and leading universities.

- Sufficient government funding to cover the cost of all aspects of the Evaluation Driven Research Paradigm.

From its inception the TIPSTER Text Program has been a jointly planned, funded, and managed program. It is unlikely that any of the individual participating Agencies could have started and sustained a program of this magnitude by itself. In addition to the three principal funding agencies, additional funds were obtained from a variety of other sources at critical junctures in the program. The most notable example of this came from the Congressionally funded Dual Use Technology Program which provided over $5 million in supplement funds in early 1992, about a quarter of the way through Phase I. This infusion of funds helped raised the TIPSTER Program to a higher level, insured that its extensive program to collect and prepare sufficient quantities of training and testing data could be completed as planned and at the desired level of quality and provided the impetus for the TIPSTER Text Program to undertake the development of its first operational prototype system based upon TIPSTER technology (i.e., the HOOKAH Project at the Drug Enforcement Administration).

◊ The implementation of the TIPSTER Phase II Architecture Demonstration System, required extensive, detailed coordination between all seven of the TIPSTER Phase II contractors. The timetable which was established for completion of this effort was extremely tight. Any single contractor who chose to drag his or her feet or not fully and openly participate would have put the completion of the whole effort in serious jeopardy. This did not happen and as a result, the TIPSTER Text Program Phase II 12-month Workshop was treated to several demonstrations of this working prototype system built in compliance with the specifications of the TIPSTER Architecture.

EVALUATION DRIVEN RESEARCH: *How Has It Performed in TIPSTER?*

Very well indeed. Following the Evaluation Driven Research Paradigm has served the TIPSTER Text Program exceedingly well. Throughout its seven year history, TIPSTER has achieved many exciting and important research results, but listing them here is beyond the intended purpose of this paper. All of the Proceedings listed in the reference section directly below are filled with excellent papers which describe in full detail what each TIPSTER Text Program participant has discovered, learned, and accomplished while investigating TIPSTER tasks under an Evaluation Driven Research Paradigm. These papers make for exciting and interesting reading and the reader is happily directed to them for further details.

SUMMARY:

During the past seven year history of the TIPSTER Text Program, there has been dramatic improvements in the current state-of-the-art in text handling, processing and exploitation. Clearly TIPSTER has been a major driving force behind these improvements within both the Information Retrieval and Information Extraction R&D communities.

While some of these advances would have happened without TIPSTER, TIPSTER was probably instrumental in accelerating their emergence. In other cases TIPSTER prodded and encouraged these R&D communities to investigate problems which they might not have considered on their own initiative.

So why has TIPSTER been able to exert such a dramatic impact over these two fields? This paper argues that this success has been made possible in large part by TIPSTER's early adoption of and continuing adherence to an Evaluation Driven Research Paradigm.

REFERENCES:

Harman D. (Ed.). *The First Text REtrieval Conference (TREC-1).* National Institute of Standards and Technology Special Publication 500-207, 1993.

Harman D. (Ed.). *The Second Text REtrieval Conference (TREC-2).* National Institute of

Standards and Technology Special Publication 500-215, 1994.

Harman D. (Ed.). *The Third Text REtrieval Conference (TREC-3).* National Institute of Standards and Technology Special Publication 500-225, 1995.

Harman D. (Ed.). *The Fourth Text REtrieval Conference (TREC-4).* National Institute of Standards and Technology Special Publication (To Appear).

Proceedings of the Third Message Understanding Conference (MUC-3), May 1991, San Francisco: Morgan Kaufmann.

Proceedings of the Fourth Message Understanding Conference (MUC-4), June 1992, San Francisco: Morgan Kaufmann.

Proceedings of the Fifth Message Understanding Conference (MUC-5), August 1993, San Francisco: Morgan Kaufmann.

Proceedings of the Sixth Message Understanding Conference (MUC-6), November 1995, San Francisco: Morgan Kaufmann (To Appear).

Proceedings of the TIPSTER Text Program (Phase I), September 1993, San Francisco: Morgan Kaufmann.

Proceedings of the TIPSTER Text Program (Phase II) , May 1996, San Francisco: Morgan Kaufmann.